Information Fusion with Constrained Equivocation

March 1998

Michael B. Hurley
MIT Lincoln Laboratory
244 Wood St., Lexington, MA 02173-9185

ABSTRACT

A review of information theory, statistical decision theory, and maximum entropy has resulted in a new technique for information fusion. Decision functions obtained from the traditional application of equivocation do not provide results that system designers will always be willing to accept; this deficiency is resolved through the addition of constraints to the equivocation function. Statistical decision theory can then be shown to be a subset of the extended equivocation theory where the constraints fully define the solution to the problem. The extended theory indicates that a rich class of decision systems exists and that decision systems can be tailored to specific applications. The performance characteristics of the decision system are specified through the constraints as opposed to the ad hoc adjustments of the cost matrix that has often been used for traditional statistical decision theory. The extended theory indicates why some approaches to information fusion have been successful while others have not. The theory also shows that consistent constraints, or system objectives, are just as important as consistent a priori knowledge for the design of distributed information fusion systems. An information fusion formula has been obtained that is valid for a subset of equivocation constraints that meet certain requirements. The formula is expected to provide good results when fusing decisions from distributed equivocation-based decision systems. Reasonable results may also be attainable when fusing decisions from systems not based upon equivocation if those systems provide reasonable performance estimates.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (FROM - TO)
01-03-1998	Conference Proceedings	xx-xx-1998 to xx-xx-1998

4. TITLE AND SUBTITLE
Information Fusion with Constrained Equivocation
Unclassified

5a. CONTRACT NUMBER
5b. GRANT NUMBER
5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Hurley, Michael B. ;

5d. PROJECT NUMBER
5e. TASK NUMBER
5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME AND ADDRESS
MIT Lincoln Laboratory
244 Wood St.
Lexington, MA02173-9185

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME AND ADDRESS
Director, CECOM RDEC
Night Vision and Electronic Sensors Directorate, Security Team
10221 Burbeck Road
Ft. Belvoir, VA22060-5806

10. SPONSOR/MONITOR'S ACRONYM(S)
11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
APUBLIC RELEASE
,

13. SUPPLEMENTARY NOTES
See Also ADM201041, 1998 IRIS Proceedings on CD-ROM.

14. ABSTRACT

A review of information theory, statistical decision theory, and maximum entropy has resulted in a new technique for information fusion. Decision functions obtained from the traditional application of equivocation do not provide results that system designers will always be willing to accept; this deficiency is resolved through the addition of constraints to the equivocation function. Statistical decision theory can then be shown to be a subset of the extended equivocation theory where the constraints fully define the solution to the problem. The extended theory indicates that a rich class of decision systems exists and that decision systems can be tailored to specific applications. The performance characteristics of the decision system are specified through the constraints as opposed to the ad hoc adjustments of the cost matrix that has often been used for traditional statistical decision theory. The extended theory indicates why some approaches to information fusion have been successful while others have not. The theory also shows that consistent constraints, or system objectives, are just as important as consistent a priori knowledge for the design of distributed information fusion systems. An information fusion formula has been obtained that is valid for a subset of equivocation constraints that meet certain requirements. The formula is expected to provide good results when fusing decisions from distributed equivocation-based decision systems. Reasonable results may also be attainable when fusing decisions from systems not based upon equivocation if those systems provide reasonable performance estimates.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	Public Release	20	Fenster, Lynn lfenster@dtic.mil
Unclassified	Unclassified	Unclassified			19b. TELEPHONE NUMBER International Area Code Area Code Telephone Number 703767-9007 DSN 427-9007

1. Introduction

Information theory naturally presents itself as a candidate applicable to classification problems because both problems can be similarly described. Information theory has traditionally been applied to communication problems which are usually modeled as a set of states transmitted through a communications channel and received as a new set of states, as shown in Figure 1. Information theory concerns itself with maximizing the information content of the transmission and minimizing the information loss that may occur.

Figure 1. An input-output communication channel

In these applications, the initial states are known, along with a priori knowledge on their relative probability of occurrence. In many applications, the initial states can be converted to a new set of states that maximize the information that is transmitted. Information theory states that the optimal distribution of input states is one where all initial states are equally probable. Usually, the information loss is not controllable, and information theory can be used to predict such things as channel capacity. Physical channels with identical information theoretic form can be compared with each other and the theoretical channel to evaluate performance.

1.1. Classification as Information Transmission

The traditional classification problem can be described as a process closely related to the description of communications systems. The problem can be formulated in the following manner, as shown in Figure 2. There is a universal space S that contains all information relevant to the system. In this space all classes in the set of classes are fully separable. This can be considered to occur by embedding the set of classes in the universe as an additional dimension. For most of the remaining discussion, only the class dimension in the universal space S is of interest, and the other dimensions will not be discussed. Measurements on the space S transform the states in the universe to the feature space V. This space is often the only space directly accessible to an observer, although in the real world, there is usually some control over the measurement process and the feature space that it creates. The classification problem is

to create a transformation or decision rule that produces predicted classes in the classification space Γ that most closely match the true classes in the universal space S.

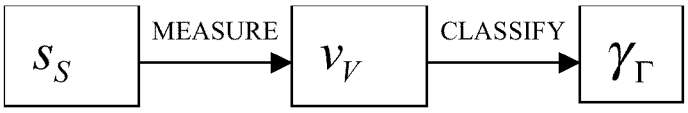

Figure 2. The classification process as a communications channel

A priori information is often available, or assumed to be available, for the design of the classifier. The a priori information usually includes the relative occurrence of each class and functions that describe the class distributions in feature space.

1.2. Statistical Decision Theory

Statistical decision theory has been used to solve the above type of problem for more than 40 years. Useful presentations of statistical decision theory are given in Ripley,[1] and in the classic text on statistical communication theory written by Middleton.[2] The approach is to make decisions that minimize some form of loss, represented by a loss function $L(k,l)$. The loss function is usually a matrix, where the elements of the matrix contain constants for the loss due to decision l if the true class is k. Generally, the matrix is defined with the diagonal terms $L(k,k)$ set to zero and the off-diagonal terms set to non-zero values. A doubt decision D is often added to the loss function with $L(k,D)=d$ for all k.

The risk function for classifier $\hat{c}$ is the expected loss when using $\hat{c}$ and is a function of the unknown class k:

$$\begin{aligned} R(\hat{c},k) &= E\left[L(k,\hat{c}(V))\,|\,C=k\right] \\ &= \sum_{l=1}^{K} L(k,l)\Pr\{\hat{c}(V)=l\,|\,C=k\} + L(k,D)\Pr\{\hat{c}(V)=D\,|\,C=k\} \\ &= \mathrm{pmc}(k) + d\,\mathrm{pd}(k). \end{aligned} \tag{1}$$

The total risk is the total expected loss, where the class C and the vector V are randomly distributed:

$$R(\hat{c}) = E\left[R(c,C)\right] = \sum_{k=1}^{K} \sigma_k\,\mathrm{pmc}(k) + d\sum_{k=1}^{K} \sigma_k\,\mathrm{pd(k)}. \tag{2}$$

This is the overall misclassification probability loss plus d times the overall doubt probability loss. The a priori probability of the occurrence of class k is σ_k. The posterior probability of class k, given $V = v$, is a conditional probability

$$p(k \mid v) = \Pr\{C = k \mid V = v\} = \frac{\sigma_k p_k(v)}{\sum_{l=1}^{K} \sigma_l p_l(v)}. \tag{3}$$

Given this, statistical decision theory delivers the following general decision rule that minimizes total risk:

$$c(v) = \begin{cases} k, \text{ where } \sum_j L(j,k) p(j \mid v) = \min_{l \le K} \sum_j L(j,l) p(j \mid v) < d \\ D, \text{otherwise} \end{cases}. \tag{4}$$

When the loss function $L(k,l)$ contains zeros on the diagonal and ones on the off-diagonals, the loss function simplifies to

$$c(v) = \begin{cases} k, \text{ where } p(k \mid v) = \max_{l \le K} p(l \mid v) > 1 - d \\ D, \text{ if each } p(k \mid v) \le 1 - d \end{cases}. \tag{5}$$

Middleton's approach to statistical decision theory relies on a function for the average loss rating

$$L(\sigma, \delta) = E_{V,S}\{\mathbf{F}(\mathrm{s}, \gamma)\} = \int_S ds \int_V dv \int_\Gamma d\gamma \mathbf{F}(\mathrm{s}, \gamma) \sigma(s) F_n(v \mid s) \delta(\gamma \mid v), \tag{6}$$

where the universal space, or signal space, is represented by S, the feature space, or observation space, by V, and the decision space by Γ. Coordinates in the three spaces are represented by s, v, and γ, respectively. The function $\sigma(s)$ represents the a priori probability of class s, $F_n(v \mid s)$ the probability distribution function for class s onto v, and $\delta(\gamma \mid v)$, the decision rule. The loss function $\mathbf{F}(\mathrm{s}, \gamma)$ is a general loss function and can take a number of forms. If the loss function is defined as a cost function that is independent of the decision function δ,

$$\mathbf{F}(\mathrm{s}, \gamma) = C(s, \gamma), \tag{7}$$

the loss function is identical to that previously described. Middleton does not specifically include the doubt class, although this approach does not rule out the possibility of different numbers of a priori and a posteriori classes. The doubt class can be ignored for the time being and will again be addressed later.

Middleton goes on to point out an additional loss function that is suggested by information theory:

$$\mathbf{F}(\mathrm{s},\gamma) = -\ln p(s \mid \gamma), \tag{8}$$

where $p(s \mid \gamma)$ is the a posteriori probability of s given γ. This loss function causes the average loss rating to be the equivocation of information theory. The average information loss is then given by

$$H(s,\delta) = E_{V,S}\{h(\mathrm{s},\delta)\} = \int_S \sigma(s)ds \int_V F_n(v \mid s)dv \int_\Gamma d\gamma[\ln p(s \mid \gamma)]\delta(\gamma \mid v). \tag{9}$$

Natural units have been adopted for the logarithm, since derivatives will be encountered later. The equivocation loss function can be shown to be a Bayesian loss function, as are the cost functions discussed earlier. It can also be demonstrated that the equivocation loss function can result in decision rules identical to cost function decision rules, but in general the decision rules are not identical. The basic conclusion of Middleton's text was that the equivocation loss function is very interesting but does not result in a reliable method for statistical decision theory, given the state of knowledge at the time that the book was written.

2. Equivocation and Information Theory

Shannon's classic work in information theory[3] resulted in a connection between information theory and statistical mechanics. Research in statistical decision theory resulted in a connection between information theory and statistical decision theory. Yu[4] provides a good introduction to information theory, although his text is intended for the application of information theory to optics. The measure of self-information in information theory is directly analogous to the entropy of a statistical system. Information I of an ensemble A is given as

$$I(A) \equiv -\sum_A P(a)\ln P(a) \equiv H(A), \tag{10}$$

where $P(a)$ is the probability of state a, and the summation is over the input ensemble A. The entropy is represented by H and is identical to Shannon's information. If the ensemble A is the input of a channel and ensemble B is the output, the self-information in ensemble B is given by

$$I(B) \equiv -\sum_B P(b)\ln P(b) \equiv H(B). \tag{11}$$

For communication theory, the entropy H is mainly a measure of uncertainty, while for statistical thermodynamics, a measure of disorder. The concept of self-information can be extended to conditional self-information:

$$I(B \mid A) \equiv -\sum_B \sum_A P(a,b)\ln P(b \mid a) \equiv H(B \mid A), \tag{12}$$

where $H(B \mid A)$ is the conditional entropy of B given A. The product ensemble AB can also be written as

$$H(AB) = -\sum_A \sum_B p(a,b) \ln p(a,b), \tag{13}$$

where $p(a,b)$ is the joint probability of both events a and b. Given the above entropy equations, the relation

$$H(AB) = H(A) + H(B \mid A), \tag{14}$$

as well as

$$H(AB) = H(B) + H(A \mid B), \tag{15}$$

where

$$H(A \mid B) = -\sum_A \sum_B p(a,b) \ln p(a \mid b), \tag{16}$$

can be derived. Average mutual information may now be defined. But first, conditional mutual information is defined to be

$$I(A;b) \equiv \sum_A p(a \mid b) I(a;b), \tag{17}$$

where

$$I(a;b) \equiv \ln \frac{p(a \mid b)}{p(a)}. \tag{18}$$

By taking the ensemble average of the above definition, the average mutual information is defined as

$$I(A;B) \equiv \sum_B p(b) I(A;b), \tag{19}$$

which can be rewritten as

$$I(A;B) \equiv \sum_B \sum_A p(a,b) \ln \frac{p(a \mid b)}{p(a)}. \tag{20}$$

From the entropy equation, it can be shown that

$$H(AB) = H(A) + H(B) - I(A;B), \tag{21}$$

$$I(A;B) = H(A) - H(A \mid B), \tag{22}$$

and

$$I(A;B) = H(B) - H(B \mid A). \tag{23}$$

These final equations are the focus of our interest, since they describe the information transfer from A to B. If $H(A)$ is the average amount information at the input of the channel, then the conditional entropy $H(A \mid B)$ is the average amount of information loss in the channel. Information theorists usually call this conditional entropy "equivocation." Others refer to it as cross-entropy, directed divergence, expected weight of divergence, or relative entropy. Also, if $H(B)$ is the average amount of information at the output of the channel, then $H(B \mid A)$ is the average amount of information needed to specify the noise disturbance in the channel and is referred to as the noise entropy of the channel.

3. Equivocation and Decision Theory

The primary use of the average mutual information equation is to determine the maximum theoretical capacity of a communications system. For decision theory, the goal is to obtain the maximum possible mutual information from a decision system. The mutual information equation most useful for decision theory is Equation (22). The entropy of the initial ensemble A is determined by the nature of the problem being solved and is often beyond the control of the decision system designer. The designer usually only has control over the equivocation. The goal of maximizing the average mutual information is replaced by the goal of minimizing the equivocation of Equation (16).

For decision theory, a set S containing an ensemble of classes describes the input data. The probability of occurrence of each member of S is given by $\sigma(s)$. The identity function for the a priori probabilities requires that they be exhaustive,

$$\sum_{S} \sigma(s) = 1. \tag{24}$$

A function $F(v \mid s)$ describes how the members of set S map into the feature space V. The identity for this probability distribution function is

$$\int_{V} F(v \mid s)dv = 1 \tag{25}$$

for each set s. These functions determine the decision rule that will minimize equivocation. The decision rule is represented by $\delta(\gamma \mid v)$, with the identity

$$\sum_{\Gamma} \delta(\gamma \mid v) = 1. \tag{26}$$

The conditional probability of class s when the decision γ has been chosen is given by

$$p(s \mid \gamma) = \frac{\int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv}{\sum_S \int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv}, \tag{27}$$

and the probability of the occurrence of decision γ is given by

$$p(\gamma) = \frac{\sum_S \int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv}{\sum_\Gamma \sum_S \int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv} = \sum_S \int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv . \tag{28}$$

The three identities for $\delta(\gamma \mid v)$, $F(v \mid s)$, and $\sigma(s)$, cause the denominator of $p(\gamma)$ to evaluate to 1. The equivocation, given the equations for $p(s \mid \gamma)$ and $p(\gamma)$, becomes

$$H(S \mid \Gamma) = -\sum_S \sum_\Gamma \int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv \ln\left(\frac{\int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv}{\sum_{S'} \int_V \sigma(s') F(v \mid s') \delta(\gamma \mid v) dv} \right). \tag{29}$$

The following substitutions are made to aid in the clarity of the differentiation procedure:

$$\begin{aligned} A_{\gamma s} &= \int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv \\ B_\gamma &= \sum_S \int_V \sigma(s) F(v \mid s) \delta(\gamma \mid v) dv = \sum_s A_{\gamma s} \end{aligned} . \tag{30}$$

The equivocation is now written as

$$H(S \mid \Gamma) = -\sum_\Gamma \sum_S A_{\gamma s} \ln\left(\frac{A_{\gamma s}}{B_\gamma} \right) \tag{31}$$

and is minimized to determine the optimal decision rule. The minimization, without specifying minimization parameters is

$$\Delta H(S \mid \Gamma) = -\sum_\Gamma \sum_S \Delta A_{\gamma s} \ln\left(\frac{A_{\gamma s}}{B_\gamma} \right). \tag{32}$$

The decision function $\delta(\gamma \mid v)$ is the only function that can be modified to minimize equivocation, since $F(v \mid s)$ and $\sigma(s)$ are a priori functions. In practice, the function $F(v \mid s)$ does yield to influence by sensor designers because the design of measurement systems is under their control. Changes in the function $\delta(\gamma \mid v)$ can be described as

$$\Delta\delta(\gamma \mid v) = \sum_k \Delta v_k \delta(v - v_k)\delta(\gamma - \gamma_{2k}) - \Delta v_k \delta(v - v_k)\delta(\gamma - \gamma_{2k-1}), \tag{33}$$

where the value Δv_k represents an incremental volume change in the combined space $V\Gamma$ from one decision γ_{2k-1} to a different decision γ_{2k}. The functions of the form $\delta(v - v_k)$ are Dirac delta functions and do not represent the decision function. The identical delta functions in v_k maintain the validity of the identity for the decision function $\delta(\gamma \mid v)$. The sum over k provides for the possibility that any modification to the decision function can be represented as a summation of incremental changes.

The change in equivocation for one change of $\delta(\gamma \mid v)$ from decision γ_1 to γ_2 for volume Δv_1 is

$$\begin{aligned}\Delta H(S \mid \Gamma) = &-\sigma(s)F(v_1 \mid s)\Delta v_1(\delta(\gamma - \gamma_2) - \delta(\gamma - \gamma_1)) \\ &\times \sum_\Gamma \sum_S \ln\left(\frac{\sigma(s)\int_V F(v \mid s)\delta(\gamma \mid v)dv}{\sum_{S'}\sigma(s')\int_V \delta(\gamma \mid v)F(v \mid s')dv}\right).\end{aligned} \tag{34}$$

Continuing the expansion, the change in equivocation becomes

$$\begin{aligned}\Delta H(S \mid \Gamma) = &-\sum_S \sigma(s)F(v_1 \mid s)\Delta v_1 \\ &\times\left(\ln\left(\frac{\sigma(s)\int_V F(v \mid s)\delta(\gamma \mid v)dv}{\sum_{S'}\sigma(s')\int_V F(v \mid s')\delta(\gamma_2 \mid v)dv}\right) - \ln\left(\frac{\sigma(s)\int_V F(v \mid s)\delta(\gamma_1 \mid v)dv}{\sum_{S'}\sigma(s')\int_V F(v \mid s')\delta(\gamma_1 \mid v)dv}\right)\right).\end{aligned} \tag{35}$$

This change shows that the decisions local to v_1 are coupled to the global decisions for the whole space. Until now, the definition of $\delta(\gamma \mid v)$ has allowed for the decision at a specific v_1 to be spread over multiple γ's. It has been proven with information theory that the optimal decision at v_1 is to assign all of

the distribution to a specific γ. This is because the function, $\sum x_i \log x_i$, is minimized by setting one x_i to 1 and the others to 0.

This result means that the value for Δv_1 is 1. The change in equivocation now becomes

$$\Delta H(S \mid \Gamma) = -\sum_S \sigma(s) F(v_1 \mid s) \times \left(\ln\left(\frac{\sigma(s)\int_{V_2} F(v \mid s)dv}{\sum_{S'} \sigma(s')\int_{V_2} F(v \mid s')dv} \right) - \ln\left(\frac{\sigma(s)\int_{V_1} F(v \mid s)dv}{\sum_{S'} \sigma(s')\int_{V_1} F(v \mid s')dv} \right) \right), \tag{36}$$

where the volume V_i represents the volume in the feature space that is assigned to decision γ_i.

Given a measurement v and an *optimal* decision function $\delta(\gamma \mid v)$ that already minimizes H, the optimal selection of γ for v is the minimum of the function

$$c(v_1) = \min_n \left(-\sum_S \sigma(s) F(v_1 \mid s) \ln\left(\frac{\sigma(s)\int_{V_n} F(v \mid s)dv}{\sum_{S'} \sigma(s')\int_{V_n} F(v \mid s')dv} \right) \right). \tag{37}$$

This function can be seen to be very similar in form to the Kullback-Leibler divergence. This fact is not surprising given that both functions have similar origins. The Kullback-Leibler divergence is used to determine the quality of fit between two parametric models. The divergence is given by the function

$$d(p, p_\theta) = \int p(x) \ln\left(\frac{p(x)}{p_\theta(x)} \right), \tag{38}$$

where $p(x)$ is a modeled density and $p_\theta(x)$ is the true density. The divergence is often interpreted as a directed distance of the modeled density from the true density. Given this interpretation, the equivocation difference formula can also be interpreted as a directed distance formula. The term $\sigma(s)F(v_1 \mid s)$, evaluated for all s, describes the distribution of classes s at v_1. This term can be interpreted as a vector in a probability space with a dimension one less than the number of classes s. The term within the logarithm describes the characteristic distribution of classes s for each γ and can be interpreted as vectors in the same space, one for each γ. The appropriate choice of γ is the one that has a class distribution closest to the class distribution s of the measurement v_1. The characteristic decision vectors also indicate the reliability of the decision rule. More reliable decision rules will have vectors with

greater separation between each other in the probability space. Less reliable decision rules will have vectors that are closer to each other.

3.1. Decision Rules Comparison

In terms of conditional probabilities and the minimization of H, the decision rule can be rewritten as

$$c(v_1) = \min_n \left(-\sum_S p(s \mid v_1) \ln\left(\frac{p(s \mid \gamma)}{p(\gamma)} \right) \right). \tag{39}$$

Statistical decision theory's general classification rule for minimization of loss is

$$c(v) = \min_\gamma \sum_S L(s,\gamma) p(s \mid v), \tag{40}$$

which means the loss from equivocation can be related to the loss matrix from statistical decision theory:

$$L(s,\gamma) + k(s) = -\ln\left(\frac{p(s \mid \gamma)}{p(\gamma)} \right), \tag{41}$$

where the constants $k(s)$ are conversion parameters that account for the probability constraints that the loss matrices do not include. Information theory implies that there is a loss or cost involved with all decisions, even correct ones. Theoretical interpretations of information theory and statistical thermodynamics support this premise.

4. The Need for Constraints

Up to this point, equivocation has looked like a very useful extension to decision theory. Unfortunately, a problem arises in that most decision systems are not concerned with minimizing information loss but with minimizing other quantities such as risk or probability of error. The types of decision systems that result from simply minimizing equivocation can have very undesirable performance characteristics. This problem can be illustrated with a simple example.

The decision system is a simple one. There are two classes with the a priori probabilities σ_1 and $1-\sigma_1$. The distribution functions from s onto v are both the uniform distribution functions from 0 to 1, $U(0,1)$. Through the symmetries of the problem, the decision process can be arranged to be the selection

of a value x in the range 0 to 1 that divides the region into two decision spaces for Γ. Through cancellation of most of the terms in the equivocation function, the final result is

$$H = -\sigma_1 \ln \sigma_1 - (1-\sigma_1)\ln(1-\sigma_1). \tag{42}$$

The equivocation does not depend on x at all. Further examination of the mutual information for this system reveals it to be zero and thus all information is lost in going from one system to the other. Any decision rule is acceptable from the equivocation perspective. Statistical decision theory, on the other hand, provides a clear rule. If the losses from both incorrect decisions are assumed to be the same, then the rule states that the selection for the whole interval from 0 to 1 should be the class with the greatest a priori probability. Equivocation therefore does not always provide desirable decision rules.

4.1. Maximum Entropy

The theory of maximum entropy illustrates a means of solving the problem. Maximum entropy has been a vigorous area of research over about the last ten years, with perhaps not quite the publicity that it deserves. The primary application of maximum entropy is to determine the best model for a given data set. The philosophy behind maximum entropy is to use the entropy or self-information equation to determine the best model, which is defined as the one that has greater entropy than any other model considered. The interpretation is that this model fits the measurement data given the a priori information while being the least informative (maximally ignorant). The maximum entropy interpretation also states that the best model is also more probable than other models.[5]

A standard technique seen in many texts on maximum entropy is the imposition of constraints upon the entropy. One example is the assignment of the probabilities $P(i \mid I)$ when the only available information is that the hypotheses are mutually exclusive and exhaustive. The sum rule results in

$$\sum_{i=1}^{m} P(i \mid I) - 1 = 0, \tag{43}$$

which has been rewritten in the form of a constraint equation. A Lagrange multiplier λ is used to expand the entropy equation through the addition of a quantity equal to zero, resulting in

$$H(A) \equiv -\sum_{i=1}^{m} P(i \mid I)\ln P(i \mid I) + \lambda\left[1 - \sum_{i=1}^{m} P(i \mid I)\right]. \tag{44}$$

The entropy is differentiated with respect to the unknown terms to find the minimum. Differentiation with respect to λ results in the sum rule. Differentiation with respect to the individual $P(i \mid I)$, and solution of the resulting system of equations results in

$$P(i \mid I) = \frac{1}{m} \tag{45}$$

and

$$\lambda = \ln(m) - 1. \tag{46}$$

The principle of maximum entropy reduces to Laplace's principle of indifference for this example.

4.2. Lagrange Multipliers

The application of constraints provides a means to design decision systems with the desired performance characteristics. Constraints may be imposed in a number of ways. However, for the remaining discussion, they will be considered as being imposed through the use of Lagrange multipliers, where the constraint equations must equate to zero. With Lagrange multipliers, the equivocation formula becomes

$$H(S \mid \Gamma) = -\sum_{S}\sum_{\Gamma}\sigma(s)\int_{V}\delta(\gamma \mid v)F(v \mid s)dv \ln\left(\frac{\sigma(s)\int_{V}\delta(\gamma \mid v)F(v \mid s)dv}{\sum_{S'}\sigma(s')\int_{V}\delta(\gamma \mid v)F(v \mid s')dv}\right) + \lambda \mathbf{g}^2, \tag{47}$$

where $\mathbf{g}$ represents the constraints on the system. Differentiation of H with respect to λ returns the constraint equation. Differentiation with respect to the other parameters results in a system of equations that must be solved. Using the previous minimization results, the resultant form is

$$\Delta H(S \mid \Gamma) = -\sum_{S}\sigma(s)F(v_1 \mid s)\left(\ln\left(\frac{\sigma(s)\int_{V_2}F(v \mid s)dv}{\sum_{S'}\sigma(s')\int_{V_2}F(v \mid s')dv}\right) - \ln\left(\frac{\sigma(s)\int_{V_1}F(v \mid s)dv}{\sum_{S'}\sigma(s')\int_{V_1}F(v \mid s')dv}\right)\right) + 2\lambda \mathbf{g}\Delta\mathbf{g}, \tag{48}$$

where the derivative of $\mathbf{g}$ depends on the formulation of the constraints. The use of the squared term in the equivocation function is so that the equivocation decision rule does not change with the adoption of constraints. When equivocation is minimized, the constraint contribution to the decision rule will be zero. The effects of the constraints on the decision rule are embodied in the integrals within the logarithmic term.

$\sum_{s=\gamma}\left(B_\gamma - \sigma_s\right)^2 = 0$	The a posteriori probabilities of γ equal the a priori probabilities of s.
$\left(\left(\sum_{\gamma \neq N} A_{\gamma N}\right) - f_{FA}\right)^2 = 0$	False-alarm constraint, given the a priori class N as the noise class.
$\left(\left(\sum_{s \neq N} A_{Ns}\right) - f_{MD}\right)^2 = 0$	Missed detection constraint, given the a priori class N as the noise class
$\sum_{s=\gamma} \nabla_{\delta(\gamma\|\nu)} A_{\gamma s} = 0$	Maximize the probability of correct decisions.
$\sum_{s=\gamma} \nabla_{\delta(\gamma\|\nu)} A_{\gamma s} L_{\gamma s} = 0$	Minimize the risk of loss.

Table 1: Possible constraints for the equivocation formula.

A sample of possible constraints is contained in Table 1. The constraints in rows 4 and 5 involve additional minimization beyond that of equivocation. The ones indicated here are identical to the constraints from statistical decision theory. These constraints fully define the decision rules so that minimization of equivocation is not required. This is an example of how statistical decision theory might be embedded into a larger theory based upon equivocation. The equivocation may still be useful with fully constrained systems to determine how much information is lost due to the constraints. Recasting the traditional statistical decision rules into the equivocation rule of Equation (37) may also prove useful in information fusion systems, as will be discussed later.

Equivocation with constraints provides the means to design a wide range of decision systems tailored to specific performance specifications. The performance requirements are directly related to the constraints imposed upon the system, whereas with traditional statistical decision theory, the performance and the system constraints are imposed indirectly upon the decision system through the loss matrix. Equivocation also allows for the selection of constraints that do not fully define the decision regions. The equivocation function provides the additional constraints that fully define the decision regions. Equivocation will define the local decision rules to match the global constraints with the least amount of information loss. The equivocation formulation allows for the design of decision systems that clearly meet the performance requirements.

5. Decision Fusion with Equivocation

Equivocation is capable of providing a mechanism for data fusion with respect to classification decisions. The formulation of the decision rule given in Equation (37) can be used to examine possible methods for data fusion. Decision fusion can be decomposed into three classes of decision systems: those with decision subsystems that operate on the same feature space, on orthogonal feature spaces, and on overlapping feature spaces. The distinction between these three classes leads to different data fusion methods. Ignorance of this distinction can result in poorly designed data fusion systems.

5.1. Decision fusion in the Same Feature Space

The first type of data fusion to consider is when multiple decision subsystems use the same feature space. In fusing multiple decisions obtained from the same features, a multitude of techniques exists, including averaging, maximum probability, and minimax methods. With a cursory examination of equivocation, it's even possible to propose schemes such as weighted-averages based upon the equivocation of each decision subsystem. The optimal solution is to determine the best decision subsystem and use it. Ideally, if multiple decision subsystems perform with different levels of success in different regions of feature space, a new, combined decision system can be created that uses the best decision subsystem for each region of feature space.

The technique of using log likelihood functions or products of probabilities is clearly incorrect in this type of fusion, because the probabilities are not independent. The correct probability product rule is of the form

$$P = p(a)p(b \mid a)p(c \mid ab)\cdots , \tag{49}$$

where the conditional probabilities have to be known to correctly do decision fusion with the product rule.

5.2. Decision Fusion in Orthogonal Feature Spaces

The second type of classifier fusion to consider is where the decision rules operate upon orthogonal feature spaces. The equivocation formula can be used by considering the original feature space V to be composed of two independent subspaces ${}^1V$ and ${}^2V$. The first assumption is that the distribution function $F(v \mid s)$ can be considered to be separable in V. Given the assumptions often implicit to data fusion, this assumption is not unreasonable. If two subspaces are not separable in $F(v \mid s)$, then some information loss will occur. The second assumption for data fusion is that $\delta(\gamma \mid v)$ is separable in V,

which is generally not the case, unfortunately. When it is separable, one of the subspaces does not contribute to the decision process and can be eliminated. We will however assume that $\delta(\gamma \mid v)$ is separable and accept the losses that arise. The decision rule after separation becomes

$$c = \min_n \left(-\sum_S \sigma(s) F_1(v_1 \mid s) F_2(v_2 \mid s) \ln \left(\frac{\sigma(s) \int_{^1V_n} F_1(v \mid s) dv \int_{^2V_n} F_2(v \mid s) dv}{\sum_{S'} \sigma(s') \int_{^1V_n} F_1(v \mid s') dv \int_{^2V_n} F_2(v \mid s') dv} \right) \right). \tag{50}$$

The fusion decision rule can be extended to as many subspaces as desired through the product rule.

$$c = \min_n \left(-\sum_S \sigma(s) \prod_j \left[F_j(v_j \mid s) \right] \ln \left(\frac{\sigma(s) \prod_j \int_{^jV_n} F_j(v \mid s) dv}{\sum_{S'} \sigma(s') \prod_j \int_{^jV_n} F_j(v \mid s') dv} \right) \right). \tag{51}$$

It is assumed that minimization of the subspaces' equivocation has occurred for the subspaces' decision rules. Comparing the equivocation of the separated spaces with the equivocation of the unified space can assess the information loss from the assumption of separable spaces. To conduct the fusion, the a priori distributions $\sigma(s)$, the values of $F_1(v \mid s)$ and $F_2(v \mid s)$, and the values of the integrals for the two subspaces over the $^jV_\gamma$ regions are needed. The fusion process can be interpreted as creating a new class composition vector for (v_1, v_2). The characteristic distribution vectors of the classes s for each γ are generated by multiplying and scaling the two sets of integrals. A new probability space is created from the two feature spaces with new locations for the characteristic distribution vectors in that space. The relative confidence in the decisions for the subspaces is embedded in the integrals over the decision regions. Less reliable decision methods will have elements more nearly equal to each other and so will have less influence on the ultimate location of the new characteristic vectors than more reliable decision methods.

With the assumption that the distribution and decision functions are separable in feature space, there are two possibilities for minimization of the resulting equivocation-like functions.[6] The first is to minimize the equivocation of the two subspaces, with the application of the appropriate constraints. The second is to minimize the decomposed equivocation function with constraints. The first method is along the lines of a decentralized fusion system; the second assumes that the central fusion system can exert control on the two subsystems. The performance characteristics of the two systems will be different. The first method leads to optimal performance at the subsystems with suboptimal performance at the fusion center. The second method leads to suboptimal performance at the subsystems in order to improve performance at the fusion center.

An additional assumption made to obtain this decision function was that the constraints imposed upon the system do not directly contribute to the decision function. In applications that adopt the constraints of statistical decision theory, the constraints may not factor out of the decision function. In that case, the designer may opt to select decision regions in feature space that come closest to the desired system performance while still retaining the above equivocation-based decision fusion formula.

5.3. Decision Fusion in Nonorthogonal Feature Spaces

It is possible to derive decision fusion rules for nonorthogonal feature spaces such as the two subspaces ${}^{1}V+{}^{C}\mathrm{V}$ and ${}^{2}V+{}^{C}V$, where ${}^{C}\mathrm{V}$ is the common subspace. However, this is a class of decision fusion that should probably be avoided if at all possible. The decisions from one subspace are conditional upon the decisions in the other subspace. Optimal fusion performance would require that the process account for the conditional dependencies between the two spaces. Most decision fusion designs would be very cumbersome if the necessary information had to be provided to obtain a high-quality fused decision rule. The decision rule would be of the general form

$$c=\min_{n}\left(-\sum_{S}\sigma(s)F_1(v_1\mid s)F_2^*(v_2\mid sv_1)\ln\left(\frac{\sigma(s)\int_{V_n}F_1(v\mid s)dv\int_{V_n}F_2^*(v\mid sV_c)dv}{\sum_{S'}\sigma(s')\int_{V_n}F_1(v\mid s')dv\int_{V_n}F_2^*(v\mid s'V_c)dv}\right)\right), \quad (52)$$

where the new function $F_2^*(v\mid sV_c)$ accounts for the conditional dependency of the second subspace upon the measurements and decisions in the first subspace.

5.4. Decision Fusion in Hybrid Systems

In more complex multisensor, multitarget environments, difficulties arise when not all the sensors observe all the targets. Thus, the decision subsystems are not able to provide decisions for the full set of targets. It can be assumed that the decision subsystems have been optimized and that the feature spaces of the subsystems do not correlate, allowing the decision rule of Equation (51) to be used. When a decision subsystem cannot provide the probability estimates for the decision fusion system, the subsystem can provide the maximum indifference solution

$$F(v_1 \mid s) = \frac{1}{N}, \quad \int_{V_n} F(v \mid s) dv = \frac{1}{N}, \tag{53}$$

where N is the number of classes. This function has the nice property that it does not change the resulting decision when fused with decisions from other subsystems. It is equivalent to an identity function. Its one drawback is that the decision values are scaled by the factor $1/N$ each time a maximally indifferent expert is added to the decision process. In a multitarget, multisensor environment where not all decision subsystems make decisions on all targets, the maximum indifference solution could be applied so that the dimensionality of the feature spaces are the same for all targets. Scaling the decision rule can eliminate the need for the inclusion of maximally indifferent experts, leading to a simpler decision rule. The reformulation of the decision fusion function gives

$$\begin{aligned} c = & \left(\sum_{s'} \sigma(s') \prod_j \left[F_j(v_j \mid s') \right] \right) \\ & \times \min_n \left(- \sum_S \frac{\sigma(s) \prod_j \left[F_j(v_j \mid s) \right]}{\sum_{s'} \sigma(s') \prod_j \left[F_j(v_j \mid s') \right]} \ln \left(\frac{\sigma(s) \prod_j \int_{V_n} F_j(v \mid s) dv}{\sum_{S'} \sigma(s') \prod_j \int_{V_n} F_j(v \mid s') dv} \right) \right), \end{aligned} \tag{54}$$

where a scale factor has been extracted outside the minimization function. Extraction of the minimization function leads to a new decision function that sacrifices the direct connection with the derivative of equivocation, but gains some desirable properties.

$$c' = \min_n \left(- \sum_S \frac{\sigma(s) \prod_j \left[F_j(v_j \mid s) \right]}{\sum_{S'} \sigma(s') \prod_j \left[F_j(v_j \mid s') \right]} \ln \left(\frac{\sigma(s) \prod_j \int_{V_n} F_j(v \mid s) dv}{\sum_{S'} \sigma(s') \prod_j \int_{V_n} F_j(v \mid s') dv} \right) \right). \tag{55}$$

The first property is that fusion with maximum indifference solutions causes no change in the value of the decision function. Because of the multiplicative nature of the fusion, the fusion of subsystem results can also be done at multiple levels. A decision fusion system with multiple targets and subsystems can be designed without accounting for subsystems not providing decisions for all targets. The preliminary fusion rule at the subsystem level is

$$F_{\Pi}(v \mid s) = \frac{\prod_{j} [F_j(v_j \mid s)]}{\sum_{s'} \prod_{j} [F_j(v_j \mid s')]}$$

$$\int_{jV_n} F_{\Pi}(v \mid s) dv = \frac{\prod_{j} \int_{jV_n} F_j(v \mid s) dv}{\sum_{S'} \prod_{j} \int_{jV_n} F_j(v \mid s') dv} \tag{56}$$

and can be applied repeatedly. A sensor with multiple decision subsystems can combine the subsystem results for transmission to the final decision system as if the results were from a single subsystem. The actual decisions are held off until the final application of the decision rule c', although subsystems are capable of independent decisions with the same decision rule.

The scaled decision rule c' has some interesting properties with respect to the types of fusion results that can be anticipated. Since smaller decision values indicate greater confidence in a decision, confidence in a decision increases if the fusion process results in smaller values. Fusion of conflicting information increases decision values and reduces confidence; complimentary information reduces the values and increases confidence. Comparison of the relative magnitudes of the decision values across the possible decisions γ provides an indication of the relative confidence in the selected decision. If all the values are the same, then all decisions are equally likely.

Interestingly enough, the decision rule can also be applied in the case of no measurements. The probability data are assumed to result from the maximum indifference solution, and the maximum of the a priori probabilities $\sigma(s)$ determines the decision. The maximum indifference case can be used to define a threshold for the final decision fusion. Minimum decision values that are greater than the maximum indifference case may be used to indicate conflicts between the subsystem's decisions. In systems where not making a decision is permissible, the appropriate action may be to delay the decision until the minimum decision value is less than the maximum indifference value. These behaviors are consistent with the assumptions that the decision subsystems are examining independent features.

6. Conclusion

The equivocation formula of information theory is applicable to decision theory when constraints are used to obtain the desired system performance. The early work in the unification of information theory and statistical decision theory failed to recognize the importance of constraints. This paper shows that all

decision systems are constructed with specific performance goals and that these goals are imposed, consciously or not, as constraints on the system design. Generally, these systems are highly nonlinear. This causes the mathematical field of nonlinear constrained optimization to be an important area of research for decision fusion.

It has been shown that decision fusion systems can be categorized by the commonality of subspaces within the feature spaces of the subsystems. The subspaces determine the types of fusion rules that may work for a given fusion system. When the decision subsystems use independent feature subspaces, a new fusion rule has been obtained. The rule possesses an identity and properties of associativity and commutativity. The decision rule takes into account the performance of the subsystems being fused together so that more reliable decision subsystems more heavily influence the final decision. The fusion rule can be used to fuse results from subsystems not designed using equivocation as long as equivalent probability distributions can be determined.

Based on the results presented in this paper, future work will concentrate on learning how best to construct decision fusion systems. Studies into optimization, distribution function characterization, feature space decomposition, conversion methods from non-information theoretic decision systems into a probabilistic framework, as well as other studies, are needed to continue to advance decision theory.

---

[1] B. D. Ripley, Pattern Recognition and Neural Networks, Cambridge, Cambridge University Press, 1996.

[2] David Middleton, An Introduction to Statistical Communication Theory (Reissue of 1960 first printing), New York, IEEE Press, 1996.

[3] C. E. Shannon, "A Mathematical Theory of Communication." Bell Syst. Tech. J., vol. 27, 379--432, 623--656, 1948.

[4]Francis T. S. Yu, Optics and Information Theory, New York: Wiley, 1976.

[5] G. L. Bretthorst, "An Introduction to Model Selection using Probability Theory as Logic," in Maximum Entropy and Bayesian Methods, Glenn R. Heidbreder, ed., Dordrecht, Kluwer Academic Publishers, 1996.

[6] Imad Youssef Hoballah and Pramod K. Varshney, "On the Design and Optimization of Distributed Signal Detection and Parameter Estimation Systems," Final Project Report RADC-TR-98-130, Rome Air Development Center, 1987.